RESEARCH PLAN PROPOSAL

Design and Development of a model for Sentiment Analysis: with special reference to user feedback on public transport services.

For registration to the degree of
Doctor of Philosophy

IN THE FACULTY OF SCIENCE

THE IIS UNIVERSITY, JAIPUR

Submitted by

Shilpa Singh
Enroll. No.: ICG/2017/24787

Under the Supervision of

Dr. Astha Pareek
Sr. Asst. Professor

Department of Computer Science

November 2018

Topic

Design and Development of a model for Sentiment Analysis: with special reference to user feedback on public transport services.

1. Introduction

Increasingly user generated thoughts has cause the lifting of opinion sites such as Twitter.com, Quora.com, and Amazon.com. These social networking and Micro-blogging sites have become a largest platform for sharing user's personal feelings, thinking of any topic or product for marketing or social liking. Sometimes, if one wants to know that the comments are positive or negative particularly in product based company where success of a company depends on opinion of different customers. These opinions can be use to analyze the user's sentiments, feelings and assessment of product. Any news article that a newspaper publishes sometimes creates negative while sometimes positive buzz about any scenario to common public. This buzz directly impacts the society on a large scale (Agarwal et al., 2016).

Traditionally, Data mining involves effective data collection and warehousing as well as computer processing.

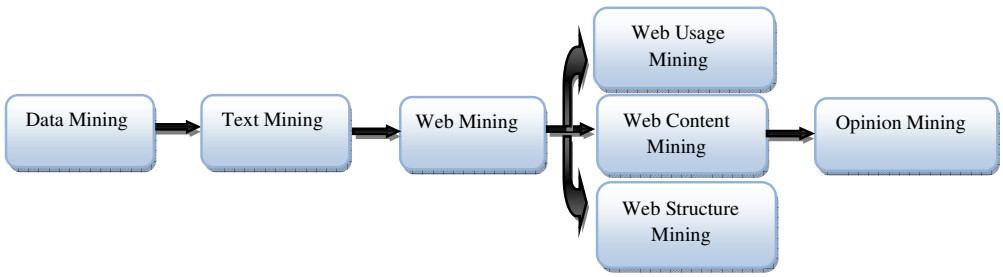

Figure 1. Hierarchy of Data Mining (Kumar and Reddy, 2016)

Figure 1 shows the different sub level of Data Mining and branches of sentiment analysis. First is the data mining, a branch of computer science is the process of extracting patterns from large data. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Next level is text mining, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers generally to the process of deriving high quality information from text. Web has become an enormous deposit of data, to which online users add new information every day. A part of that information is represented by online reviews. Extracting opinions from user generated reviews online known as web mining. Web mining is divided into three parts Web Usage Mining, Web Content Mining and Web Structure Mining. Next level is opinion mining which comes under Web content mining.

Opinion mining or sentiment analysis is a type of natural language processing (NLP) for tracking the attitudes, feelings or appraisal of the public about particular topic, product or services (Salanova et al., 2011). It is an automated extraction of subjective content from text and identifying the orientation such as positive, negative or neutral in that text. It aims to explore feelings of a person who write the text. It uses natural language processing and machine learning

ethics to determine opinion in the text (Salanova et al., 2011). “Sentiment analysis is the computational study of people's opinions, sentiments, emotions, and attitudes. This fascinating problem is increasingly important in business and society. It offers numerous research challenges but promises insight useful to anyone interested in opinion analysis and social media analysis.” (Liu, 2015).

Sentiments can be classified at various levels: Aspects or feature level, Sentence level and Document level. Aspects or feature level sentiment classification classifies the sentiments based on the sentiments polarity of each aspects or feature about some target object and sentence level sentiment classification on the other hand classifies each sentence based on their sentiment polarity towards some topic. In document level sentiment classification the polarity of whole document is determined (Rahmath and Ahmad, 2014).

1. Sentiment analysis techniques

Sentiment analysis techniques are divided into two parts first is machine learning based technique and other is lexicon based as shown in figure 2.

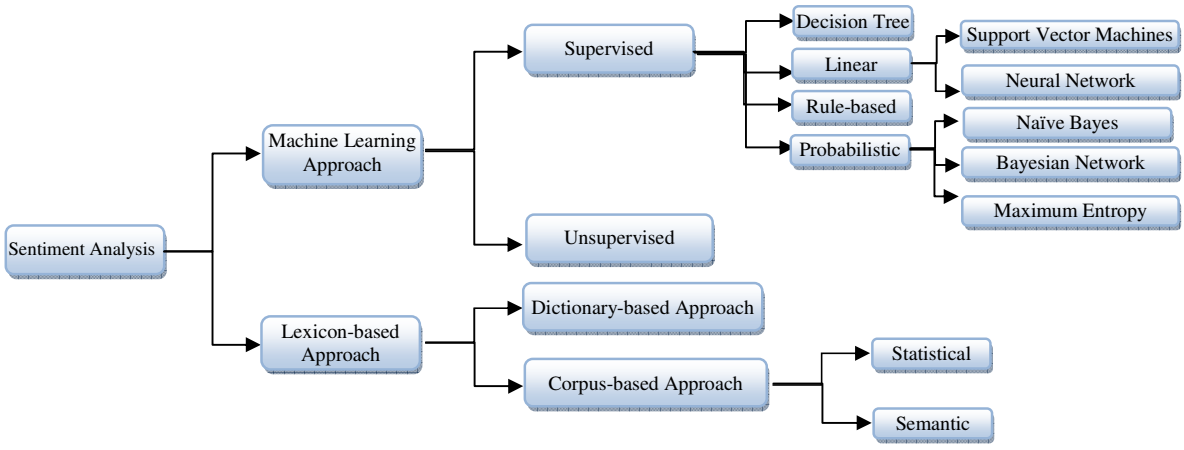

Figure 2. Sentiment classification techniques (Medhat et al., 2014)

Some researchers combined the supervised machine learning and lexicon based approaches together to improve sentiment classification performance which is called hybrid technique. Hybrid technique can be use for finding sentiment analysis on public transport services. In this technique first lexicon based corpus method will be used to find the polarity of any sentence as positive, negative and neutral. After finding the polarity machine learning approach like Naïve Bayes and SVM will be use to find accuracy.

1.1 Lexicon based technique: Lexicon based method is an unsupervised learning approach because it does not require prior training data sets. In Lexicon based technique, a sentiment dictionary with sentiment words are used for sentiment classification. The dictionary contains polarity of each word whether they are positive, negative and objective words.

1.2 Machine learning based techniques: The machine learning approach belongs to supervised classification approach. This approach is more accurate because each of the classifiers is trained on a collection of representative data.

1.2.1 Naive Bayes Classification: A Naive Bayes classifier is a simple probabilistic classifier based on Bayes theorem and is particularly suited when the size of training set is less. Naive Bayes is optimal for certain problem classes with highly dependent features.

1.2.2 Support Vector Machines: Support vector machines (SVMs) have been shown to be highly effective at traditional text categorisation. They are large-margin, rather than probabilistic, classifiers. The basic idea behind the training procedure is to find a maximum margin hyper plane, represented by vector that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible.

Other Sentiment analysis techniques are:

I. **Maximum Entropy:** Maximum Entropy (ME) classification is yet another technique, which has proven effective in a number of natural language processing applications. Maximum Entropy provides better performance when conditional independence assumptions are not met.

II. **Bayesian classification:** Bayesian classification is based on Bayes Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

The development and application of sentiment analysis required the use of powerful software tools. As the quality of accessible tools keeps on developing the decision of the most suitable tool turns out to be progressively difficult. Every tool has its own benefits and drawback. Tool like Natural Language Toolkit (NLTK) is a library in python that provide a base building programs and classification of data. It is resource for python that can be used for text processing, classification, tagging and tokenization. This toolbox plays a key role in transforming the data in the tweets into a format that can be used to extract sentiment from them. NLTK support various machine learning algorithm which are used for training classifier and to calculate accuracy of different classifier. It provides over 50 corpora and lexical resources along with the WordNet integration (Goel et al., 2016).

Other Sentiment analysis tools are:

I. **Rapid Miner Tool:** Rapid miner tool is open source licensed software for sentiment analysis. It encompasses data analysis, data integration and reporting in a single suit. It also provides important features like data preprocessing, visualization, statistical analysis, evaluation, deployment and predicting analysis. It is very easy to use software with lots of features like cross validation, performance vector, split validation and so on.

II. **WEKA Tool:** Weka is free and open source software that have a large range of opinion mining and sentiment analysis algorithms. Weka have many algorithms like classification, data preprocessing, regression, visualization, clustering and association can be use in combination or in isolation.

One can find the opinion of any person online with the help of these tools and techniques in various areas such as movies, tourism, agriculture, analysis for news and blogs, exploiting

social relations, determine public opinions, government schemes, cab services and education shopping etc. Lots of research work has been done on the various areas like Movies, Tourism and many more but it was observed that lack of research work has been done on public transport services for finding sentiment analysis of user. Our core focus is on finding Sentiment analysis in public transport services because it play an important role in maintaining the sustainability of urban transportation systems, especially in cities where buying and maintaining a car is expensive. Public transports provide better accessibility where daily commuting in cities is tough job. So, sentiment analysis can be use to analysis the opinion of users regarding the ride or for driver.

2. Review of Literature

Numerous research works have been carried out towards sentiment analysis. The sentiment analysis or opinion mining is the study of opinions, sentiments, beliefs and attitudes that people express about different topics. This analysis involves tasks as detection, extraction and classification of sentiments (Ramírez et al., 2018). The sentiment analysis also contain some important fields which includes Emotion Detection (ED), Building Resources (BR) and Transfer Learning (TL). Emotion detection aims to extract and analyze emotions, while the emotions could be explicit or implicit in the sentences. Transfer learning or Cross-Domain classification is concerned with analyzing data from one domain and then using the results in a target domain. Building Resources aims at creating lexica, corpora in which opinion expressions are annotated according to their polarity, and sometimes dictionaries (Soni and Patel, 2014). In order to understand our work better, some relevant works about several research field related to our interest will be introduces.

M D et al., (2016) compared the various techniques used for Sentiment Analysis by analyzing various methodologies. She has given a comparative table of three approaches used in sentiment analysis that is Machine Learning, Rule Based and Lexicon Based Approach on the bases of Classification, Advantages and Disadvantages. She also compared the advantages and disadvantages of various machine learning methods.

Techniques used in sentiment analysis can be machine learning or lexicon based. In machine learning we have supervised or unsupervised machine learning. (BiswaRanjanSamal et al., 2017) reflected the performance analysis of supervised machine learning techniques of sentiment analysis. For this analysis author collected the datasets of different sizes such as 10600, 25000, 35600, 50000 and 85600 on movie review. This dataset is tested and trained using NLTK package toolkit. After training of dataset sentiment analysis of movies review has been calculated using seven supervised machine learning algorithms. These seven techniques are as follows Naive Bayes, Multinomial Naive Bayes, Bernoulli Naïve Bayes, Logistic Regression, Stochastic Gradient Descent, Linear SVM/ Linear SVC and Nu SVM/Nu SVC. The result obtained concludes linear SVC/SVM as the best classifier among others in achieving 100% accuracy for large number of movie review. Author used supervised machine learning further work can be done for investigating effectiveness considering big datasets using the unsupervised and semi supervised machine learning techniques.

Vyas and Uma, (2018) performed an experiment on Rapid miner tool to derive sentiments from tweets. Author has used different algorithm to find accuracies and then compared to find the best performing one. 450 tweets on customer feedbacks have been taken from twitter. This data is splits into two parts one is training and other is testing with the split ratio =.66 it means 2/3 of data is used for training and rest of the data is used for testing. SVM, decision tree and naive

bayes techniques are used to find accuracy, precision and Recall. Accuracy of these three models has been observed with respect to different sample size of training dataset. Author used only three techniques, experiment can also be done using other techniques in Rapid miner tool.

Vishal Vyas used rapid miner with SVM, decision tree and naïve bayes but(Goel et al., 2016) used SentiWordNet along with Naive Bayes to improve accuracy of classification of tweets, by providing positivity, negativity and objectivity score of words present in tweets. The dataset consist of 1.6 million tweets of 140 dataset for different movies. This dataset has been taken for training data. After training system, when tested with a small testing dataset of 100 tweets the system gives an efficiency of 58.40%. Author concludes that by use of sentiwordnet along with Naive Bayes for classification this accuracy can improved to a considerable extent. Experiment done by author ankur goel takes a lot of time in fetching data from twitter and in data management and this flaw can be overcome by use of Hodoop ecosystem instead of python.

Some author used machine learning algorithm with python language like (Gupta et al., 2017) gives the methodologies for sentiment analysis using machine learning algorithm on Python. Data is collected on the topics Commerce, politics and Sports Events form twitter using Tweepy and Twitter API. After collecting the dataset preprocessing of data has been done using Python standard library. This dataset is divided into two sets training and testing with the ratio 70% for training and 30% for testing. Finally average accuracies of different models like SVM, Naïve Bayes, DAN2, Netural Network etc. has been calculated using Python with the result of 85% to 90% accuracy. Further research can be done with other topics and for the short form of words.

Python language can also be used in lexicon language for finding sentiment analysis. (Agarwal et al., 2016) analyzed and tested an algorithm in which news headlines can be classified as positive or negative sentiments using NLTK tool of python language. An experiment has been done by author on top 10 world news headlines of 3rd Dec 2015. In this experiment author has been calculated the positive or negative polarity of each headline manually as well as from experimental algorithm using Sentiwordnet. The result has been calculated for finding deviation from experimental result to manual result. If any error detected it was marked as E (error) otherwise they were marked C (Correct Result). Finally error percentage has been calculated. The above algorithm run for around 500 news headlines of 30 days and analyzed the deviation of experimented values from expected values of polarity. This experiment found average Deviation of 30 days news headlines was 2.7. For a huge data, this manual classification task can be erroneous and a tedious job and can deviate the results. For solving this problem we can classify them using SVM technique or apply some other machine learning task for the same.

While clustering document using lexicon chain one can face four problems. These problems are disambiguating the polysemous and synonymous words, overcoming high dimensionality, determining the number of clusters, and assigning appropriate description for the generated clusters.. Author (Wei et al., 2015) described these four problems in document clustering. Also identified and described a methodology for clustering using WordNet and Lexical chains. A modified WordNet-based semantic similarity measure has been proposed for word sense disambiguation, and lexical chains has been employed to extract core semantic features that express the topic of documents. The author solved four problems in document clustering using hybrid method. Author use the reuters-21578 corpus which has been widely used for evaluating document clustering algorithms. Cluster quality has been evaluated by three metrics that is purity, F1-measure, and entropy. Methods used are Base, DC, DCS, ASG03, LMJ10, CSF11. He concluded that the lexical chain features (core semantics) can improve the quality significantly with a reduced number of features in the document clustering process. Some limitation of using

this technique will be important words which are not included in WordNet lexicon will not be considered as concepts for similarity evaluation.

Jagdale et al., (2016) reflected the use of TwitterR package to access the live tweets from twitter API using R tool. Tweets have been accessed over time with different topics like #Budget2016, #RailBudget2016, #Freedom251, #MakeInIndia, etc. After collection and pre-processing of data sentiment score has been calculated from the events using opinion lexicons and scoring function. Finally Wordcloud has been plot for #MakeInIndia which display frequent word of event on plot and highlight those terms which helps easily on which users has given tress. Further work can be done using different tool and also Wordcloud can be plot for huge dataset.

Sentiment analysis has various methodologies and their advantages and disadvantages as described by (MD et al., 2016). These techniques can be divided into machine learning or lexicon based. In machine learning we have supervised or unsupervised machine learning. (BiswaRanjanSamal et al., 2017) reflected the performance analysis of Supervised machine learning techniques like Naive Bayes, SVM, Logistic Regression etc using NLTK package toolkit. But (Goel et al., 2016) used lexicon based technique with Naïve Bayes for finding sentiment analysis. Sentiment analysis can be done in Rapid miner tool for finding accuracy, precision, Recall using SVM, decision tree and naïve bayes techniques as explained by (Vyas and Uma, 2018). Some author used machine learning algorithm with python language like (Gupta et al., 2017) and some author used lexicon language with python for finding sentiment analysis. So from this literature review we can find out various tools and techniques used for sentiment analysis.

3. Motivation

Public transport services play an important role in maintaining the sustainability of urban transportation systems, especially in cities where buying and maintaining a car is expensive. Also, in India vast number of people from residential areas uses these facilities. However people have lots of mixed opinion about these services some are satisfied and some are not satisfied with the services like price, driver, safety, cleanliness etc. provided by the public transport such as cabs, metros, airlines, railways etc. In this way the prime target is to finding the opinion of customers of these services and to enhance the nature of these services by providing the feedback. This motivates us to select these services as a starting problem and applying that sentiment analysis technique to our research work.

Public transport service is the challenging area of research as very few researches are being carried out for finding sentiment analysis from customer feedback. Literature review reflects that several works have been done in machine learning and lexicon based technique but there is less work done in hybrid technique in which first lexicon based corpus method is used to find the polarity of any sentence and then machine learning approaches like Naïve Bayes, SVN etc is used to find accuracy. This is also observed in recent research that NLTK is the most popular toolkit used by the authors but this tool is not used for finding sentiment analysis of public transport services.

This motivates us to enhance the literature by working on it. The present work derived from sentiment analysis may help these services to identify the cause i.e. main reason of being unsatisfied customers, so the corrective measures may be taken at appropriate level at suitable time.

4. Objectives

The major objectives set for the research work are:

1. To study literature on existing Sentiment Analysis techniques on the available public transport services.
2. To collect and classify the reviews based on different services and categories from social media.
3. To develop a model by enhancing existing techniques.
4. To evaluate the new model by analyzing users feedback on public transport services in identified region and group of people.

5. Methodology

The research will adopt a methodology as the main focus is to achieve its objectives. Firstly detailed review of existing literature in sentiment analysis will perform to identify state-of-art in algorithms, methodology and problem domains. Method will be evaluated for the collection and storage of data using NLTK toolkit. This data is first split into sentences then every sentence will classify as positive negative or neutral with the help of sentiwordnet. Different techniques will be tie up into a generic implementable framework for the analysis of user reviews. Linguistic approach will be adopted by utilizing the sentiment properties in language structures. Model and schema will be developed for demonstrating classification outcomes. Experiment will be conducted as a means of evaluating the usefulness of the extracted information and the accuracy of classifications. Conclusions will be drawn on the basis of the above experimentation. Lastly discussion will be done leading to final conclusions and scope for future work.

6. Plan of work

6.1. Work Plan:

Phase 1 (4-5 months):

Literature survey of Sentiment analysis, Techniques, public transport Services, NLTK tool, Sentiwordnet and Python Language etc.

Phase 2 (2-3 months):

Detail study of NLTK tool, Lexicon based methods with Machine learning methods etc.

Phase 3 (6-7 months):

Investigation of available public transport services data in twitter. Finding the polarity of data as positive, negative and objective using wordnet and sentiwordnet etc.

Proposed improvements / new algorithm to find sentiment analysis.

Phase 4 (4-10 months):

Validation of improvement/ new algorithm.

Phase 5 (4-5 months):

Analyze the result obtained and thesis writing.

6.2. Gantt chart:

Project Plan	Months											
	2	4	6	8	10	12	14	16	18	20	22	24
Phase 1												
Phase 2												
Phase 3												
Phase 4												
Phase 5												

Figure 3 Gantt chart

7. Place of work and facilities available

To follow the line of investigation the research will be carried out in the laboratories existing in the university. The hardware and software available here will be used as a part of the research work.

8. References, bibliography and webliography

Agarwal, A., V. Sharma, G. Sikka and R. Dhir (2016). Opinion Mining of News Headlines using SentiWordNet. *IEEE*. **16**, 978-1-5090-0669-4.

Balaji, H., V. Govindasamy and V. Akila (2016). Social Opinion Mining and Concise Rendition. *ICACCCT*. 978-1-4673-9545-8.

BiswaRanjanSamal, A. K. Behera and M. Panda (2017). Performance Analysis of Supervised Machine Learning Techniques for Sentiment Analysis. *IEEE*. 978-1-5090-4929-5.

Goel, A., J. Gautam and S. Kumar (2016). Real Time Sentiment Analysis of Tweets Using Naive Bayes**.** *IEEE.* **16**, 978-1-5090-3257-0.

Gupta, B., M. Negi, K. Vishwakarma, G. Rawat and P. Badhani (2017). Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python. *International Journal of Computer Applications.* **165**, 0975 – 8887.

Hridoy, S. A. A., M. T. Ekram, M. S. Islam, F. Ahmed and R. M. Rahman (2015). Localized twitter opinion mining using sentiment analysis. *SpringerOpen Journal.* **1186**, 40165-015-0016-4.

Hughes, G. and C. Dobbins (2015). The utilization of data analysis techniques in predicting student performance in massive open online courses. *SpringerOpen Journal.* **1186,** s41039-015-0007-z.

Jagdale, R. S., V. S. Shirsat and S. N. Deshmukh (2016). Sentiment Analysis of Events from Twitter Using Open Source Tool. *IJCSMC*. **5**, 475-485.

Ji, Y., Y. Cao, Y. Du and H. M. Zhang (2017). Comparative Analyses of Taxi Operations at the Airport. *ELSEVIER.* **25**, 2222-2232.

Kumar, B. S. and D.B. S. Reddy (2016). An Analysis on Opinion Mining: Techniques and Tools. *PARIPEX - INDIAN JOURNAL OF RESEARCH*. 5, 2250-1991.

Liu, B. (2015). *Sentiment Analysis, Mining Opinions, Sentiments, and Emotions*, p 381, Cambridge University Press, Chicago.

M D, D., S. C and A. Ganesh (2016). Sentiment Analysis:A Comparative Study On Different Approaches. *ELSEVIER*. **87**, 44-49.

Medagoda, N., S. Shanmuganathan and J. Whalley (2015). Sentiment Lexicon Construction Using SentiWordNet 3.0. *IEEE*. **15.** 978-1-4673-7679-2.

Medhat, W., A. Hassan and H. Korashy (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*. **5**, 1093-1113.

Mishra, N. and C. K. Jha (2012). Classification of Opinion Mining Techniques. *International Journal of Computer Applications.* **56**, 0975-8887.

Patil, H. P. and M. Atique (2015). Sentiment Analysis for Social Media: A Survey. *IEEE*. **15,** 978-1-4673-8611-1.

Rahmath, P H. and T. Ahmad (2014). Sentiment analysis Techniques – A Comparative Study. *International Journal of Computational Engineering & Management*. **17**, 2230-7893.

Ramírez, F. J., G. Alor, J. L. Sánchez, B. A. Olivares and L. Rodríguez (2018). A Brief Review on the Use of Sentiment Analysis Approaches in Social Networks. *Springer International Publishing*. **1007,** 978-3-319-69341-5_24.

Salanova, J. M., M. Estrada, G. Aifadopoulou and E. Mitsakis (2011). A review of the modeling of taxi services. *ELSEVIER*. **20**, 150-161.

Shukla, R., A. Chandra and H. Jain (2017). OLA VS UBER: The Battle of Dominance. *IOSR*. *2319-7668.*

Soni, V. and M. R Patel (2014). Unsupervised Opinion Mining From Text Reviews Using SentiWordNet. *International Journal of Computer Trends and Technology*. **11,** 2231-2803.

Terrana, D., A. Augello and G. Pilato (2014). Automatic Unsupervised Polarity Detection on a Twitter Data Stream. *IEEE*. **14**, 978-1-4799-4003-5.

Vyas, V. and V. Uma (2018). An Extensive study of Sentiment Analysis tools and Binary Classification of tweets using Rapid Miner. *ELSEVIER.***125,** 329-335.

Wei, T., Y. Lu, H. Chang, Q. Zhou and X. Bao (2015). A semantic approach for text clustering using WordNet and lexical chains. *ELSEVIER*. **42**, 2264-2275.